Wrangle Report

In this project wrangle WeRateDogs Twitter data were used to create interesting and trustworthy analyses and visualizations. In addition to this data, there were two additional data types from different sources to complete the analysis in a good way.

This report describe what have been done to achieve this goal.

Project details:

This project pathes through three main steps:

- Data Gathering
- Data Assessment
- Data Cleaning

1- Data Gathering

Three data types have been gathered:

1- **Enhanced Twitter Archive, this contain:**

Tweet's text, rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo). This is *.csv file.

2- **Twitter API, this contain:**

Tweet ID, retweet count, and favorite count. This is *.txt file.

3- **Image Predictions File, this contain:**

A table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images). This is *.tsv file.

2- Data Assessment

The assessment has been done through:

- Visual assessment: in Excel file and in Jupyter notebook by printing the inter data.
- Programmatically: by using .info, .duplicated, .sample

The problems occurred were as following:

Tidy Problems;

- non descriptive columns (p1, p1_conf, p1_dog, etc)
- Redundant Column for prediction

Quality Problems;

- Missing values (NaN)
- timestamp is object insted of integer
- retweeted_status_timestamp coulmn is object instead of integer
- none values in dog stage
- name coulmn has inappropriate values (a, an , the)
- denumerator has no. less than 10 , make it 10
- redundant columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id)
- duplicated jpg url
- tweet id int and has to be obj

3- **Data Assessment**

It passes also by three steps:

- Define the problem and identify how to solve it
- Write the code in order to solve these problems
- Test the code that has been done